El Proceso de Extracción de Conocimiento en la Determinación del Perfil del Autor y la Atribución de Autoría

Mercado V.[1], Villagra A.[2], Errecalde M.[3]

[1-2] Laboratorio de Tecnologías Emergentes (LabTEm), Instituto de Tecnología Aplicada (ITA)
Unidad Académica Caleta Olivia, Universidad Nacional de la Patagonia Austral.
Santa Cruz - Argentina.

[3] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC),
Departamento de Informática - Universidad Nacional de San Luis. San Luis - Argentina.

{vmercado,avillagra}@uaco.unpa.edu.ar, merreca@unsl.edu.ar

RESUMEN

En el presente trabajo se describen, brevemente, las tareas de investigación y desarrollo que se están llevando a cabo en forma conjunta en el área de análisis de autor de documentos entre el LIDIC de la UNSL y el LabTEm de la UNPA. En particular, se ha tomado como caso de estudio primario los documentos de periodistas con diversas orientaciones políticas (oficialista vs opositor) con el objetivo de realizar con los mismos el Análisis de Autor y la Determinación/Caracterización del perfil del autor. Ambos tipos de tareas, han ganado creciente interés en la comunidad científica internacional y en empresas dedicadas al análisis de la información en la Web, por lo que la línea de investigación propuesta permitiría la formación de

académico / científico como en la industria.

Palabras claves: Minería de Textos, Análisis de Autoría, Determinación del perfil del Autor, orientación política en Artículos Periodísticos. El proceso KDD.

CONTEXTO

Esta línea de trabajo se enmarca en los trabajos conjuntos que desde hace varios años llevan a cabo investigadores del LabTEm de la UNPA y el LIDIC de la UNSL. En particular, las tareas de investigación desarrolladas tienden a consolidar trabajos previos conjuntos relacionados a la Minería de Textos y la Web [Taquias et al., 2014], y complementarlos con los desarrollos que en el LIDIC se están llevando a cabo en las áreas específicas de análisis de autoría y determinación del perfil del autor [Funez et al., 2013., Villegas et al., 2014].

En este contexto, ambos laboratorios no sólo disponen de financiación obtenida de proyectos de investigación además se fluidas de investigación con centros de excelencia mundial especializados en estos temas como el Laboratorio de Tecnologías del Lenguaje del INAOE (Puebla, México) y el *Artificial Intelligence Laboratory*-DICSE de la *University of the Aegean* (Karlovassi, Grecia). En particular, una integrante del LabTEm desarrollará su trabajo de Maestría en esta temática, mientras que en el LIDIC un becario de

doctorado y uno post-doctoral de CONICET trabajarán en la temática específica de determinación del perfil del autor, y colaborarán en aquellos temas que se solapen con la presente investigación.

1. INTRODUCCIÓN

A partir de la disponibilidad de volúmenes inmensos de información en la Web, se reconoce cada día más el rol de la Minería de Datos (MD) como una herramienta fundamental para hacer un uso adecuado y ventajoso de esta información. Esta tendencia crece día a día y se plantean nuevos escenarios relevantes como es el caso de *Big Data*, donde el contexto en el cual deben ser aplicados los métodos de MD es sumamente desafiante. En particular, un área que comienza a ganar creciente interés es la *determinación del perfil del autor* (DPA), es decir, aquella que identifica patrones compartidos por un *grupo de gente* y que aborda problemas de clasificación de los usuarios de la Web de acuerdo a la edad, género, orientación política, etc. La DPA, un sub-campo del área más general conocida como *análisis de autoría* (AA), es un tema muy importante de investigación por sus potenciales (y actuales) aplicaciones en problemas de seguridad nacional e inteligencia, lingüística forense, análisis de mercados e identificación de rasgos de personalidad, entre otros. Otro sub-campo de la AA muy estudiado, denominado *atribución de autoría* (ATA), consiste en la atribución de un texto de autoría desconocida a uno de un conjunto de autores potenciales.

Si bien la MD, la DPA y la ATA son áreas de investigación científica muy activas, cuando se aplican a problemas concretos de la vida real se las debe considerar en el contexto más general del proceso de extracción de conocimiento, que involucra varias etapas y herramientas para la recopilación de información, pre-procesamiento y extracción de características, análisis y visualización. El problema es que, usualmente, estas herramientas están dispersas, escritas en lenguajes y plataformas diferentes y, en muchos casos, como en el análisis de información textual, no están disponibles para el idioma español.

En este contexto, esta línea de investigación se propone el abordaje de dos tareas de AA, una de ATA y otra de DPA, como lo son la atribución de autoría y la determinación de la orientación política en documentos periodísticos, en el contexto de un proceso completo de extracción de conocimiento.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

En esta sección se describen las líneas de investigación que se llevan a cabo en el proyecto:

El *análisis de autoría* (AA) [Stamatatos, 2009] es un área de investigación que ha ganado interés creciente en los últimos años principalmente por sus potenciales (y actuales) aplicaciones en problemas de seguridad nacional e inteligencia, lingüística forense, análisis de mercados e identificación de rasgos de personalidad, entre otros. El AA se enfoca en la clasificación automática de textos basándose fundamentalmente en las elecciones estilísticas de los autores de los documentos, e incluye distintas tareas de análisis como, por ejemplo: a) la *atribución de autoría*, b) la *verificación de autor*, c) la *detección de plagios*, d) la *determinación del perfil del autor* y e) la *detección de inconsistencias estilísticas*.

Los enfoques predominantes en esta área están basados en el aprendizaje automático/de máquina supervisado. En pocas palabras, estos enfoques derivan, a partir de un conjunto de datos etiquetados (conjunto de entrenamiento) y un proceso inductivo de aprendizaje/entrenamiento, un clasificador que puede generalizar sus predicciones a otros datos no observados previamente. La representación clásica de los textos/documentos en estos casos, incluye tanto atributos basados en el contenido (palabras) como en el estilo de escritura de los autores.

A partir de la disponibilidad de volúmenes inmensos de información en la Web, se reconoce cada día más el rol de la AA como una herramienta fundamental para hacer un uso adecuado y ventajoso de esta información, lo que ha quedado plasmado en un incremento de Workshops y Competencias específicos de esta temática. En particular, un área que comienza a ganar creciente interés es la determinación del perfil del autor, es decir, aquella que identifica patrones compartidos por un grupo de gente y que aborda problemas de clasificación de acuerdo a la edad y género [Peersman et al., 2011, Schler et al., 2006, Argamon et al., 2009], nacionalidad, personalidad [Celli et al., 2014, Mairesse et al., 2007], orientación política [Abooraig et al.,2014, Conover et al.,2011, Malouf & Mullen, 2007], etc.

Más allá de la relevancia y ventajas que pueden tener este tipo de tareas existe, actualmente, un desarrollo limitado en nuestro país de trabajos y grupos de investigación especializados en la problemática del AA. En este contexto, en esta línea de investigación nos enfocaremos en dos áreas claves de la AA como lo son la *determinación del perfil del autor* (DPA), y la *atribución de autoría* (ATA).

Respecto a la DPA, también conocida como *caracterización del autor* (en inglés *author profiling*), incluye actividades como la determinación automática de la edad, género, rasgos de personalidad y orientación política, entre otras. En nuestro caso, nos concentraremos en la orientación política (pro-gobierno vs opositor) de documentos periodísticos de acceso público, como libros de investigación periodística, blogs periodísticos, artículos en revistas y diarios on-line, etc.

Respecto a la ATA, analizaremos las particularidades que surgen para la identificación automática de autores, en aquellos contextos en donde los mismos tienen igual o diferente orientación política. En estos casos, se analizará cuáles son las *features* (estilográficas o de contenido) que son más relevantes para discriminar los distintos autores que pertenecen al mismo (o diferente) espectro político.

Por otra parte, a diferencia de los estudios de laboratorio, donde es usual disponer de datos recolectados y procesados *a priori*, listos para ser analizados, el proceso de extracción de conocimiento (en inglés *KDD*, por *Knowledge Discovery in Data*) [Kurgan & Musilek, 2006, Fayyad et.al., 1996] involucrado en problemas prácticos concretos requiere de varias etapas y herramientas para la recopilación de información, pre-procesamiento y extracción de características, análisis y visualización. El problema es que, usualmente, estas herramientas están dispersas, escritas en lenguajes y plataformas diferentes y, en muchos casos, como en el análisis de información textual, no están disponibles para el idioma español. Si bien existen hoy en día nuevas herramientas y plataformas como

KNIME[1] y RapidMiner[2] que se suponen asisten al usuario en identificar e integrar estas etapas y herramientas, no siempre es claro cómo compatibilizan estas plataformas aspectos como la claridad, flexibilidad, facilidad de uso y extensión, entre otros. Por lo tanto, realizar una experiencia concreta sobre uno o varios problemas particulares (como la DPA y la ATA) utilizando una plataforma de este tipo, permitirá ganar experiencia que podrá servir no sólo en problemas de Minería de Textos y de la Web sino en otras tareas de análisis futuros que involucran otros datos arbitrarios como, por ejemplo, imágenes, videos, sonido, datos de redes de sensores, etc.

3. RESULTADOS OBTENIDOS / ESPERADOS

En cuanto a los resultados se pretende lograr un sistema integrado de *atribución de autoría* con periodistas de la Argentina y de *determinación de la orientación política en documentos periodísticos*, que también soporte el *descubrimiento de tópicos* en estos documentos. Este resultado quedará plasmado en distintos *workflows* del tipo de los soportados por KNIME, en los cuales los distintos "nodos" que componen las tareas quedan explícitamente expresados, facilitándose su uso y modificación por parte de aquellos usuarios no familiarizados con este tipo de tareas. Estos *workflows* contarán además con nodos dedicados a la evaluación y clara visualización de los resultados obtenidos.

El sistema anterior será utilizado en trabajos experimentales realizándose comparaciones con enfoques similares representativos del estado del arte en el área. Un objetivo adicional a largo plazo es que la experiencia obtenida con estas tareas sirva para abordar otros procesos de KDD que involucran otros tipos de datos (imágenes, videos, etc.) lográndose así consolidar en la UNPA un equipo de trabajo especializado en temáticas de gran relevancia nacional e internacional.

[1] *https://www.knime.org/*

[2] *https://rapidminer.com/*

4. FORMACIÓN DE RECURSOS HUMANOS

Un integrante de este proyecto de investigación cuenta con una beca Post-Doctoral de CONICET enfocada en aspectos psicológicos relacionados a las tareas de DPA.

Un integrante de este proyecto de investigación está desarrollando su Tesis de Doctorado sobre DPA multimodal con una beca Doctoral de CONICET.

Un integrante está desarrollando su tesis de Maestría orientada en esta línea de investigación.

5. REFERENCIAS

[1] **Abooraig** R., Alwajeeh A., Al-Ayyoub M., and Hmeidi I.**(2014)**. On the automatic categorization of arabic articles based on their political orientation. In *Proc. of the Third International Conference on Informatics Engineering and Information Science (ICIEIS2014)*.

[2] **Argamon S**., Dhawle S., Koppel M. and Pennebaker J. (**2005**). *Lexical Predictors of Personality Type*. Joint Annual Meeting of the Interface and the Classification Society of North America.

[3] **Celli** F., Lepri B., Biel J.-I., Gatica-Perez D., Riccardi G., and Pianesi F.**(2014)**. *The workshop on computational personality recognition* 2014. In Proceedings of the ACM International

Conference on Multimedia, MM '14, pages 1245-1246, New York, NY, USA. ACM.

[4] **Conover** M., Goncalves B., Ratkiewicz J., Flammini A., and Menczer F.**(2011)**. *Predicting the political alignment of twitter users*. In Proceedings of 3rd IEEE Conference on Social Computing (SocialCom).

[5] **Kurgan** L. A. and Musilek P. (**2006**). *A survey of Knowledge Discovery and Data Mining process models*. Knowledge Engineering Review. 21, 1 (March 2006), 1-24**.**

[6] **Fayyad** U. M., Piatetsky-Shapiro, G. and Smyth, P. (**1996**). *From data mining to knowledge discovery: an overview*. In Advances in knowledge discovery and data mining, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (Eds.). American Association for Artificial Intelligence, Menlo Park, CA, USA 1-34.

[7] Funez D. G., Cagnina L., and Errecalde M. L.**(2013)**. *Determinación de género y edad en blogs en español mediante enfoques basados en perfil*. In Anales del XIX Congreso Argentino de Ciencias de la Computación (CACIC *2013),* pages 1003-1012.

[8] **Mairesse** F., Walker M. A., Mehl M. R. and Moore R. K. (**2007**). *Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text*. In JAIR, 30, 457-500.

[9] **Malouf** R. and Mullen T.**(2007)** *Graph-based user classification for informal online political discourse*.

[10] **Peersman** C., Daelemans, W. and Van Vaerenbergh L. **(2011)**. *Predicting age and gender in online social networks.* In Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11, pages 37-44, New York, NY, USA. ACM.

[11] **Schler** J., Koppel M., Argamon S., and Pennebaker, J. W. **(2006)**. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs,* pages 199-205, 2006.

[12] **Taquias** D., Villagra A., and Errecalde M. L.. Detección de plagios con adversarios. In Anales del XVI Workshop de Investigadores en Ciencias de la Computación (WICC 2014), pages 233, 237, 2014.

[13] **Villegas** M. P., Garciarena-Ucelay M. J., Errecalde M. L. and Cagnina L. C. (**2014**). *A Spanish Text Corpus for the Author Profiling Task*, XX Congreso Argentino de Ciencias de la Computación, Buenos Aires, Argentina.